DOCUMENT RESUME

ED 377 227 TM 022 402

AUTHOR Bergstrom, Betty A.; And Others
TITLE Differential Item Functioning vs Differential Test Functioning.
PUB DATE Apr 93
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Error of Measurement; Grade 11; Grade 12; High Schools; High School Students; *Item Bias; Item Response Theory; Mathematics; *Sample Size; Sampling
IDENTIFIERS Calibration; Michigan Educational Assessment Program

ABSTRACT

A problem that arises when a differential item functioning (DIF) study is done with samples of examinees differing in ability is examined. A test may function differently when the populations from which the items are calibrated are not of equal ability. Since the lower ability examinees get many difficult items incorrect, the spread (standard deviation) of item calibrations may differ. The difference in how the test functions, in terms of item spread, must be addressed before differential item function can be explored. Several methods to account for differences in standard deviation are reported. The problem is studied using data from the Michigan Educational Assessment Program (MEAP) for 619 eleventh and twelfth graders from the 1991 administration of the MEAP mathematics test. Either linear transformation or subsampling helps eliminate differences in standard deviation, but subsampling is preferable when there is a large enough sample. Five figures and three tables illustrate the analysis, along with eight figures in an appendix describing sample groups. (Contains 9 references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BETTY A. BERGSTROM

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ED 377 227

# Differential Item Functioning vs Differential Test Functioning

Betty A. Bergstrom
Richard C. Gershon

Computer Adaptive Technologies, Inc.

William L. Brown
Minneapolis Public Schools

Paper presented at the annual meeting
of the American Education Research Association

April, 1993
Atlanta, Georgia

To receive additional copies please contact either of the first two authors at Computer Adaptive Technologies, 2609 E. Lunt Avenue-Suite2, Chicago, IL 60645 (312) 274-3286.

TM022402



BEST COPY AVAILABLE

Abstract

We examine a problem which arises when a differential item functioning (DIF) study is done with samples of examinees differing in ability. A test may function differently when the populations from which the items are calibrated are not of equal ability. Since the lower ability examinees get many difficult items incorrect, the spread (standard deviation) of item calibrations may differ. The difference in how the test functions, in terms of item spread, must be addressed before differential item function can be explored. We report on several methods to account for differences in standard deviation.

Differential Item Functioning vs
Differential Test Functioning

There is a problem which arises when a differential item functioning (DIF) study is done with samples of examinees differing in ability. A test may function differently when the populations from which the items are calibrated are not of equal ability. Since the lower ability examinees get many difficult items incorrect, the spread (standard deviation) of item calibrations may differ. The difference in how the test functions, in terms of item spread, must be addressed before differential item function can be explored.

The Carl D. Perkins Vocational and Applied Technology Act Amendment of 1990 requires State Departments of Education to develop accountability systems that document the academic progress of vocational education students (Merkel-Keller, 1992). To meet these federal requirements, the State of Michigan is considering using an existing criterion referenced test, the Michigan Educational Assessment Program (MEAP), to assess progress for vocational education students. The MEAP was designed to measure essential skills in mathematics; reading and science (Michigan Department of Education, 1990). By state mandate, the MEAP is administered each year to all 4th, 7th and 10th grade students. In Spring, 1992, a pilot sample of vocational education students took sections of the MEAP (Gershon and Bergstrom, 1992). The goal of the project was to explore the use of the 10th grade MEAP as a pre-test. At a later point, a different form of the MEAP would be administered to assess academic progress. It would be advantageous to use the existing MEAP because of the time and cost involved in administering additional tests to students.

We compare item calibrations obtained from students enrolled in vocational education programs and students drawn from the general population, to determine whether items on the *Conceptualization and Problem Solving* section of the MEAP function differentially for the two groups.

The vocational education sample was lower in ability than the general population sample. We also explore the impact that difference in ability across groups has on DIF and outline the procedures we followed to account for the resulting disturbance in the DIF study.

## Method and Results

### Samples

In the Fall semester, the MEAP is administered to all 10th grade students in the state of Michigan. The state draws an official random sample for research purposes. This sample of 2,040 students from the 1991 test administration of the MEAP Essential Skills Mathematics Test was used in our study.

Five Intermediate School Districts (ISD) from throughout Michigan participated in a research project to administer sections of the MEAP to 11th and 12th grade students in vocational education programs. These tests were administered to 619 students in Spring, 1992, by school personnel at the ISD vocational education centers.

### Test Specifications

A subsection of 50 items from the *Conceptualization and Problem Solving* section of the Essential Skills Mathematics Test was used for our analysis. The content objectives covered in this section are: 1) whole numbers and numeration, 2) fractions, decimals, ratio and percent, 3) measurement, 4) geometry, and 5) problem solving and logical reasoning (Michigan State Board of Education, 1989).

Comparison of Item Calibrations

We analyzed the data with the Rasch model software BIGSTEPS (Wright and Linacre, 1991). Separate item calibrations were obtained from the general population sample and the vocational education sample. The separate calibration *t*-test approach:

$$t_i = \frac{d_{i1} - d_{i2}}{\sqrt{s_{i1}^2 + s_{i2}^2}}$$

where $d_{i1}$ is the difficulty of item $i$ in the calibration based on the general population sample, $d_{i2}$ is the difficulty of item $i$ in the calibration based on the vocational education sample, $s_{i1}$ is the standard error of estimate for $d_{i1}$ and $s_{i2}$ is the standard error of estimate for $d_{i2}$ was used to detect differences in calibrated difficulty of items between the two calibrations (Wright and Stone, 1979; Smith, 1992). If items are functioning similarly for both samples, the estimates of their difficulties should be statistically equivalent and the value of $t$ within the range $-1.96 < t < 1.96$. This method defines DIF as a statistically significant difference in the calibrated difficulty of the item for the two samples.

Table 1 shows the descriptive statistics from the two calibrations. Item difficulties were mean centered (mean= 0.00). But, note the difference in the standard deviation for the two item calibrations. Items calibrated by the general sample (SD = .93) were more widely distributed than items calibrated by the vocational education sample (SD = .76). Table 1 also shows that vocational education students (mean = -0.44 ) performed less well than the general population sample (mean = -0.19).

In Figure 1, the two sets of item calibrations are plotted against each other. While some items appear to be functioning differentially, the line obtained by regressing the calibrations obtained from the vocational education sample on the calibrations obtained from the general population sample shifts approximately 10 degrees from the identity line. Many

TABLE 1
DESCRIPTIVE STATISTICS

| | N | Mean Ability Estimate | SD of Ability Estimate | Mean Standard Error |
|---|---|---|---|---|
| Item Calibrations | | | | |
| General Sample | 50 | 0.00* | .93 | .05 |
| Vocational Sample | 50 | 0.00* | .76 | .09 |
| Persons | | | | |
| General Sample | 2040 | -0.19 | .90 | .33 |
| Vocational Sample | 619 | -0.44 | .89 | .32 |

*value assigned

of the items found outside the 95% confidence interval lines are located at the ends of the distribution. We suspected that the shift of the regression line away from the identity line was due to the difference in the standard deviation of the two sets of calibrations and that the standard deviation difference was caused by a difference in the ability levels of the two samples.

TABLE 2
GENERAL POPULATION SAMPLE-ABILITY GROUPS

| Group | N | Percent Correct | Mean Ability Estimate | SD of Ability | SD of Items |
|---|---|---|---|---|---|
| 1 | 71 | ≤ .20 | -1.67 | .24 | .07 |
| 2 | 321 | > .20 and ≤ .30 | -1.11 | .15 | .62 |
| 3 | 463 | > .30 and ≤ .40 | -.66 | .13 | .80 |
| 4 | 432 | > .40 and ≤ .50 | -.20 | .14 | 1.00 |
| 5 | 338 | > .50 and ≤ .60 | .27 | .15 | 1.20 |
| 6 | 215 | > .60 and ≤ .70 | .86 | .17 | 1.36 |
| 7 | 110 | > .70 and ≤ .80 | 1.54 | .19 | 1.52 |
| 8 | 67 | > .80 | 2.42 | .57 | 1.38 |

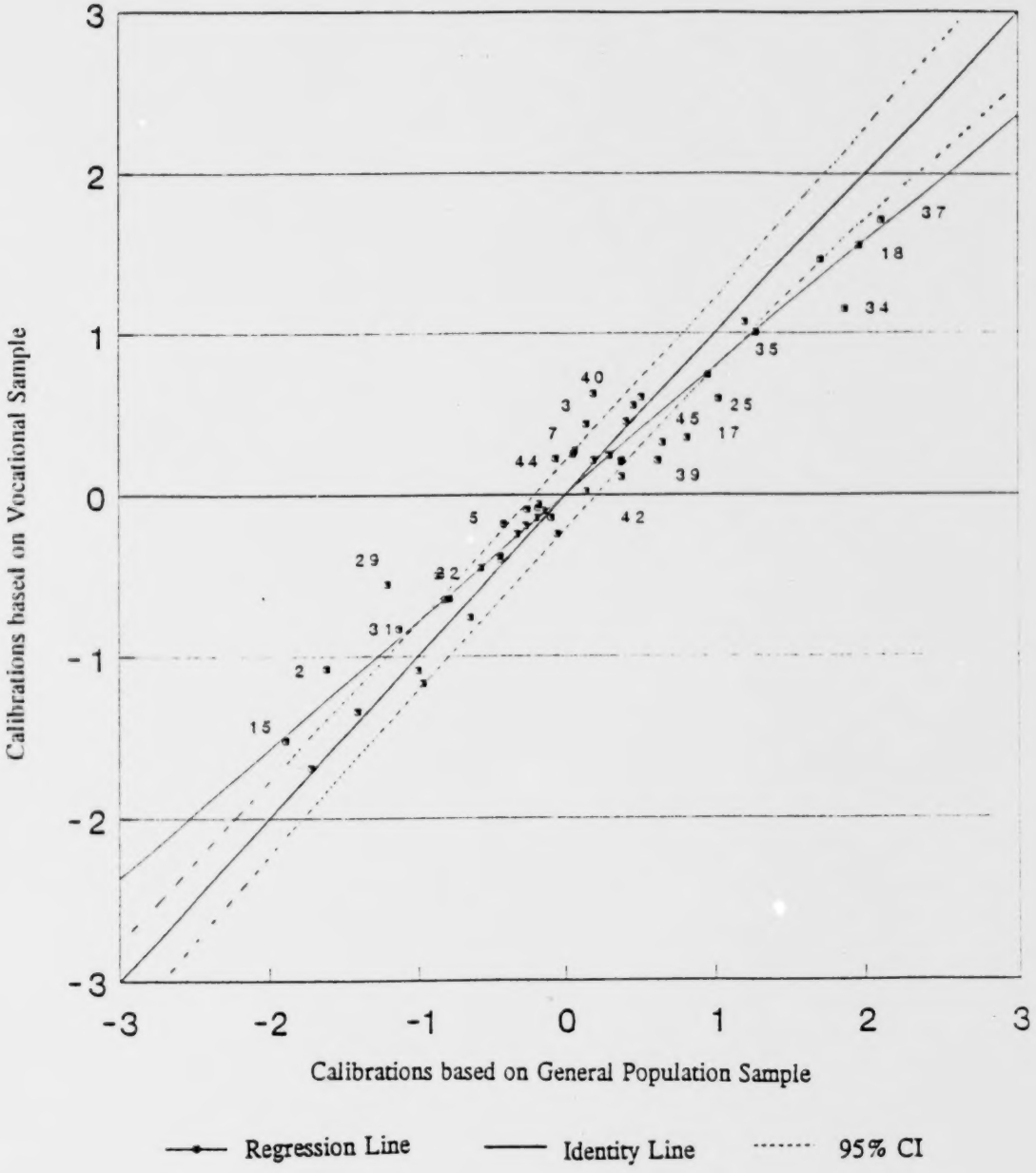



Figure 1. Comparison of Item Calibrations

The Effect of Ability on the Standard Deviation of Item Calibrations

We used the general population data to explore the effect of differing ability level groups on standard deviation. We divided the examinees into eight groups based on the percentage of items they correctly answered. Table 2 shows the mean ability in logits for each group. We then ran a separate analysis for each group. Figure 2 shows how the standard deviation of the item calibrations varies systematically with the ability of the sample used for calibration. Low ability groups separate the items less than high ability groups. When the percentage of items correct is greater than 80%, the standard deviation of the item calibration shrinks, implying that very high ability groups don't separate the items at the easy end of the scale. This change in how the test functions can be followed by examining the BIGSTEPS *Maps of Persons and Items* for each group (See Appendix).

Adjusting for Differences in the Standard Deviation

In order to adjust for differences in the standard deviations of the two sets of item calibrations, we transformed the calibrations derived from the vocational education sample using the following linear transformation:

$$d'_{i2} = d_{i2} * ( SD_1 / SD_2 )$$

where $d_{i2}$ is the difficulty of item $i$ in the calibration based on the vocational education sample, $SD_1$ is the standard deviation of the calibrations based on the general population sample and $SD_2$ is the standard deviation of $d_{i2}$.

Figure 3 shows the comparison of the two sets of calibrations after stretching the distribution of the calibrations derived from the vocational education sample such that the standard deviations of the two sets of calibrations were equal (Mean = 0.00, SD = .93).

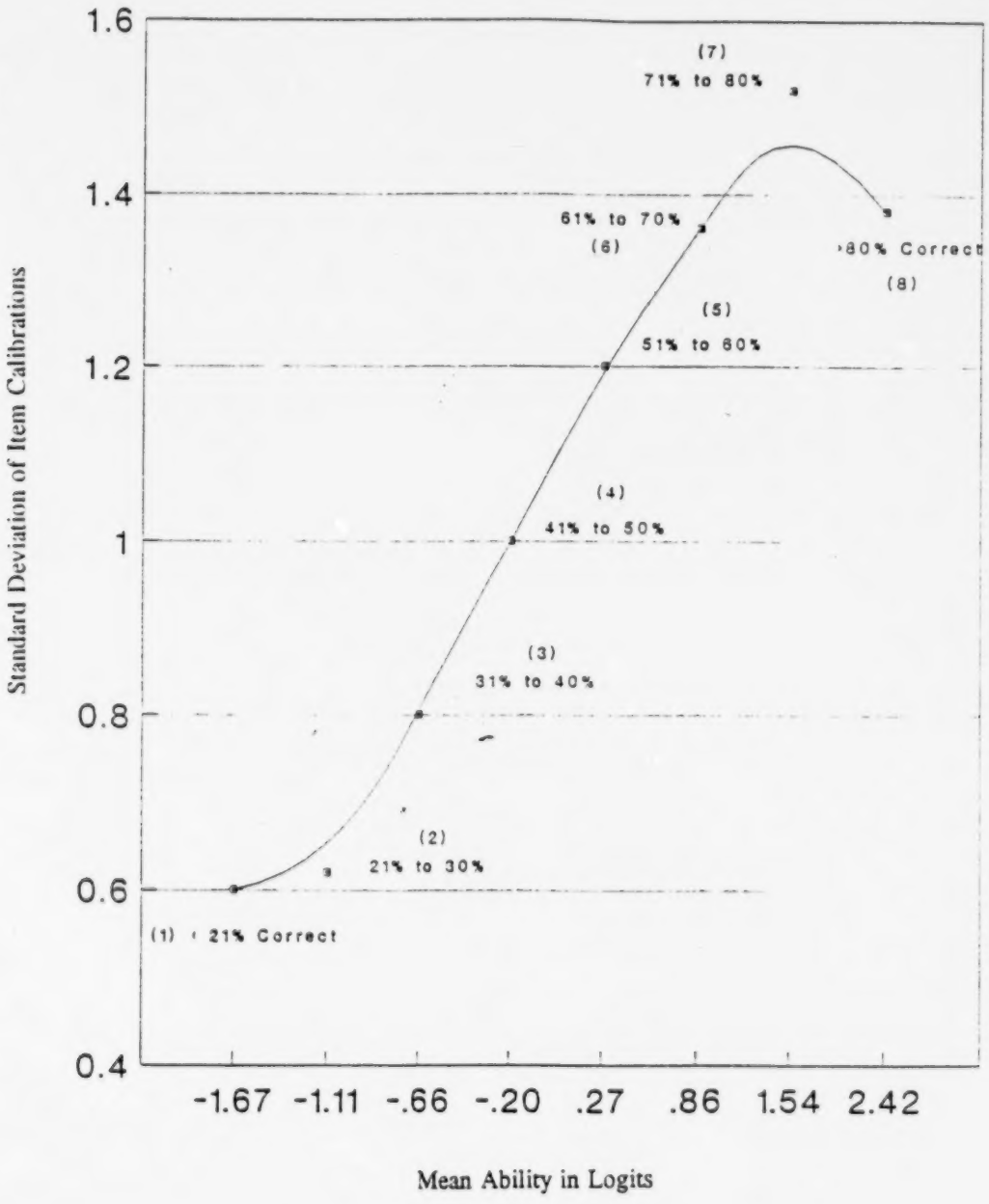



Figure 2. Standard deviation of item calibrations by mean ability groups

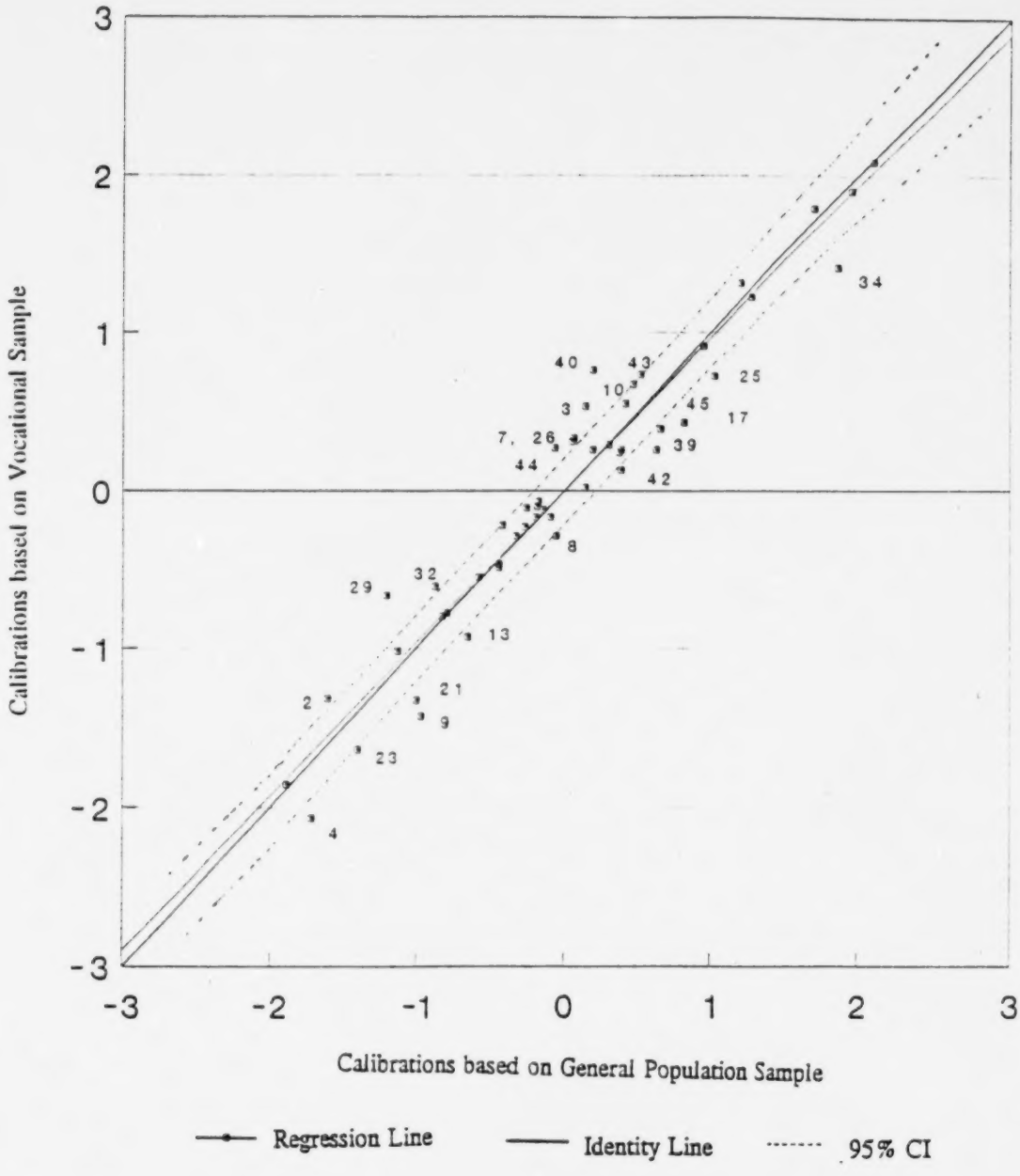



Figure 3. Comparison of item calibrations after using a linear transformation to adjust for differences in standard deviation.

Now the calibrations can be examined for DIF, free from the confounding influence of difference in item difficulty standard deviation caused by ability level.

Figure 3 shows that 10 items had a $t$ value $< -1.96$, indicating that they were significantly more difficult for the vocational education students, while 12 items had a $t$ value $> 1.96$ indicating that they were significantly easier for the vocational education students. While some of the items identified in Figure 1 as functioning differentially for the vocational education students still appear in Figure 3, some items such as items 15, 18 and 37 are no longer identified as functioning differentially when the difference in standard deviations is eliminated.

Comparing Samples of Equal Ability

Another way of thinking about DIF is that, when items function differentially, the probability of a correct response for persons of equal ability, but different group membership, is not the same (Scheuneman, 1991). A second method for accounting for differences in the standard deviation of item calibrations is to subsample from each sample and compare how items functioned for examinees of approximately equal ability. Figure 4 shows the distribution of the vocational education sample and the general population by the percentage of items they got correct on the test.

We drew a subsample from each of the samples, including only those examinees who answered $> 40\%$ or $< 70\%$ of the items correctly. These are the examinees for whom the most information would have been obtained from each item (Wright and Stone, 1979). Table 3 shows the descriptive statistics for the subsample analyses. The item difficulties are mean centered and the ratio of the item calibration standard deviations is 1.02, compared with 1.22 from Table 1. Student ability estimates are also comparable.

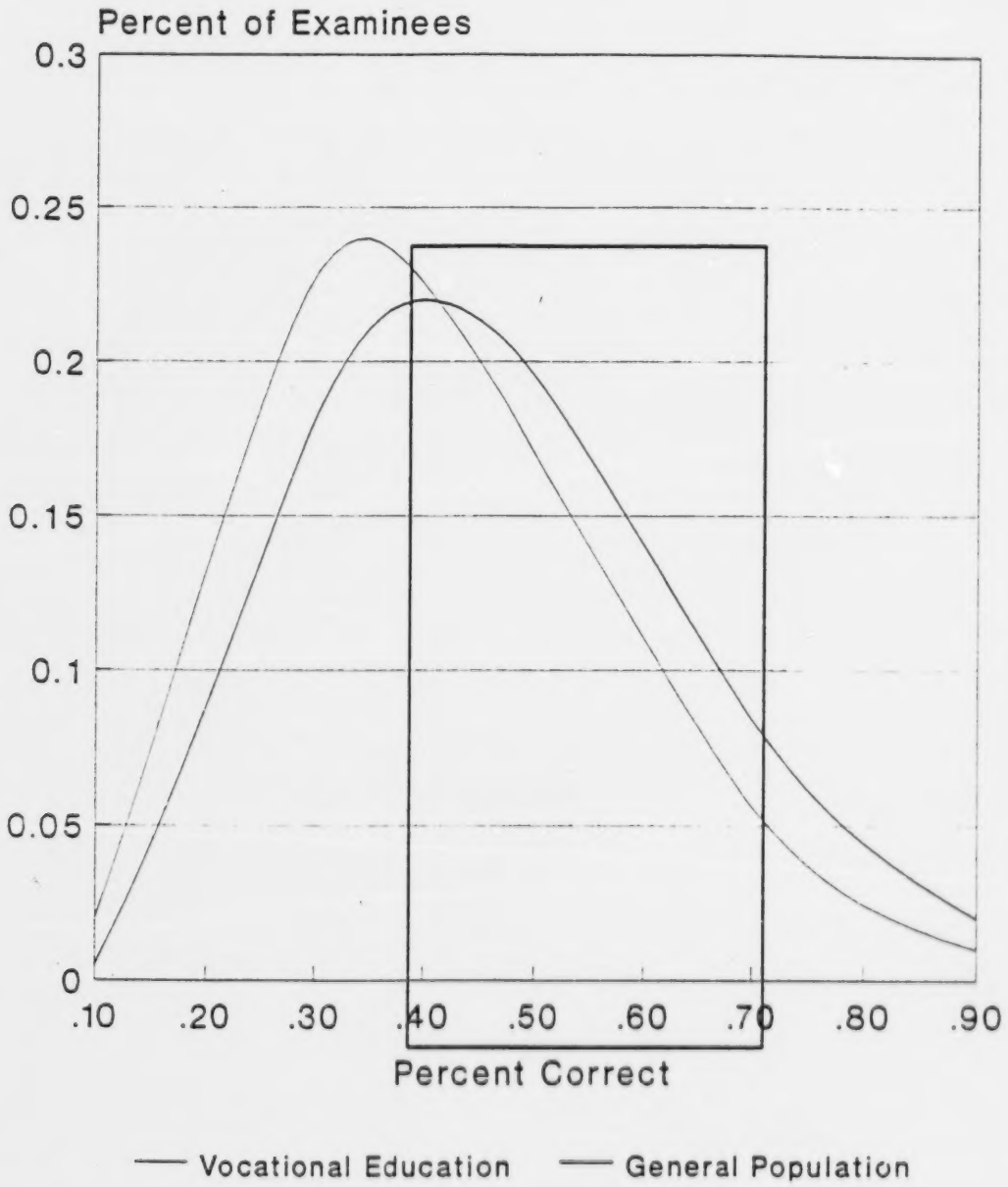



Figure 4. Comparison of Ability Distributions by Percent Correct

TABLE 3
DESCRIPTIVE STATISTICS--SUBSAMPLES

| | N | Mean Ability Estimate | SD of Ability Estimate | Mean Standard Error |
|---|---|---|---|---|
| Items | | | | |
| General Sample | 50 | 0.00 | 1.12 | .07 |
| Vocational Sample | 50 | 0.00 | 1.10 | .15 |
| Persons | | | | |
| General Sample | 985 | 0.17 | .42 | .32 |
| Vocational Sample | 250 | 0.14 | .40 | .32 |

Figure 5 shows how the items functioned when only those examinees for whom the test is appropriately targeted are included in the analysis. Using 1.96 as the *t*-value cutoff point, 4 items are significantly easier for vocational education students while 6 items are significantly more difficult for vocational education students.

Comparison of Methods

The correlation for the *t*-values obtained using the linear transformation and the subsampling method was .91. The subsampling method, however, reduced the sample size, increasing the standard error of measurement for the item calibrations (refer to Tables 1 and 3) and thus decreasing the power to detect statistically significant DIF (see Smith, 1993 for additional discussion on power). With one exception, both methods identified the same items as having the greatest difference in calibration between the general population and the vocational education students. Item 17 had a *t*-value of 3.71 using the linear transformation but only 1.66 using the subsampling method. This item was difficult (calibration = 1.07 for the subsample general population) and the fit statistics indicated that this item data may have been spoiled by guessing.

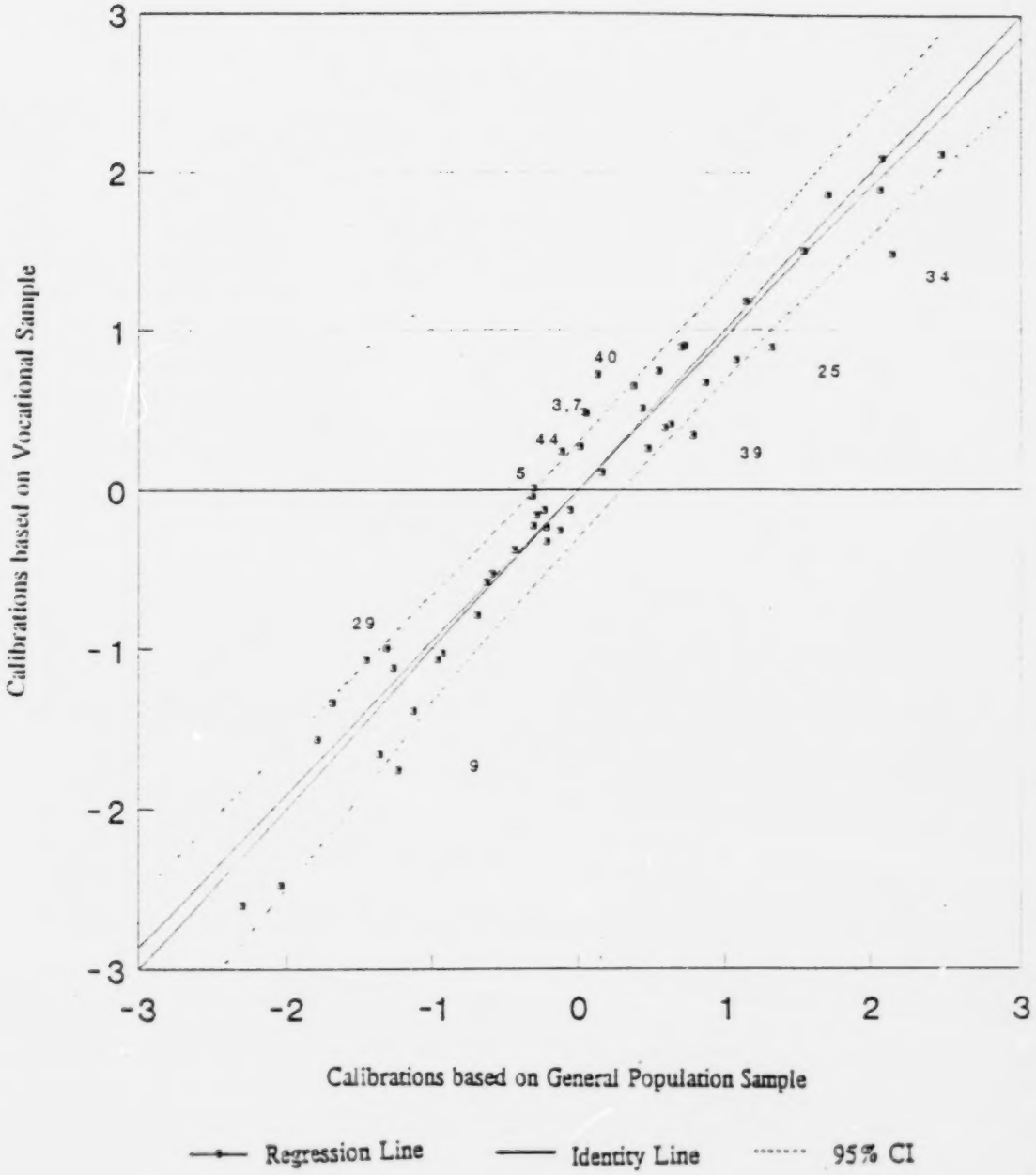



Figure 5. Comparison of Item Calibrations derived from selected subsamples

Deleting inappropriate person-item interactions

We tried an additional method for removing differences in the standard deviation of the item calibrations. Using the CUTHI, CUTLO options in BIGSTEPS, we marked as missing data, examinee/item encounters where the examinee's ability minus the item difficulty was less than -1 or was greater than 2 (Gershon, 1992). This bases item calibration only on responses from examinees of appropriate ability. When an item is too difficult or too easy for an examinee, the response is treated as missing. This method however, failed to reduce the difference in the standard deviations of the item calibration because the vocational education sample was still of lower ability than the general population sample. Greater restrictions using CUTHI and CUTLO produced very small vocational education sample sizes for difficult items.

Item Content

When we examined the items which were functioning differentially, we found that, to some extent, items which required abstract reasoning were more difficult for vocational education students while items which had concrete examples and/or required visual-spatial logic were easier for vocational education students.

## Discussion

Either linear transformation or subsampling works to eliminate differences in standard deviation. Subsampling is preferable when you have a large enough sample because it eliminates problems introduced when items are taken by examinees for whom the item is too difficult or too easy.

The Rasch model specifies that raw score is the sufficient statistic for measurement. This means that when constructing a test from an item bank, it must not make a difference



BEST COPY AVAILABLE

which items happen to be included in the test. If items show DIF, the raw score is no longer a sufficient statistic (Linacre, 1992). This is especially important when different forms of a test are used to measure gain across time. Interaction between gain and item type is a possibility when items are known to function differentially for the subpopulation for which the gain score is computed, and the proportion of item types in the test changes from form to form. For example, in this application, if the post test contains 15% more items based on concrete examples than the pre-test, and vocational education students appear to improve over time, is the "improvement" due to actual gain, or due to an increased percentage of items which these students find easier?

## Conclusion

Everyone wants to avoid repetitious or unnecessary testing. Exploring the use of an existing instrument to meet federal requirements is a sensible procedure. However, studies must be undertaken to determine item types contained in a test, differential item functioning across item types, and the proportion of item types across test forms before an instrument can be used to measure gain. When differential item functioning is studied, differences in ability between the subpopulation of interest and the total population must be taken into consideration.

References

Gershon, R.C. (1992, April). The effect of person-item mismatches on the integrity of the item characteristic curve. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

Gershon, R.C. and Bergstrom, B.A. (1992). *Final report: Standard 1 pilot project*. Lansing, Michigan. Department of Education.

Linacre, J.M. (1992,Fall). Why fuss about statistical sufficiency? *Rasch Measurement*, 6,3 230.

Merkel-Keller, C. (1992, April). *Hands on assessment: Beyond the pencil and paper in vocational education*. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

Michigan Department of Education (1990, September). *Michigan Educational Assessment Handbook*.

Michigan State Board of Education (1989, March). An interpretation of the Michigan Essential goals and objectives for mathematics education. Lansing, Michigan.

Smith, R.M. (In Press) *Applications of Rasch Measurement* Chicago: MESA Press.

Wright, B.D. and Linacre, J.M. (1991) *BIGSTEPS* (Computer Program) Chicago: MESA Press.

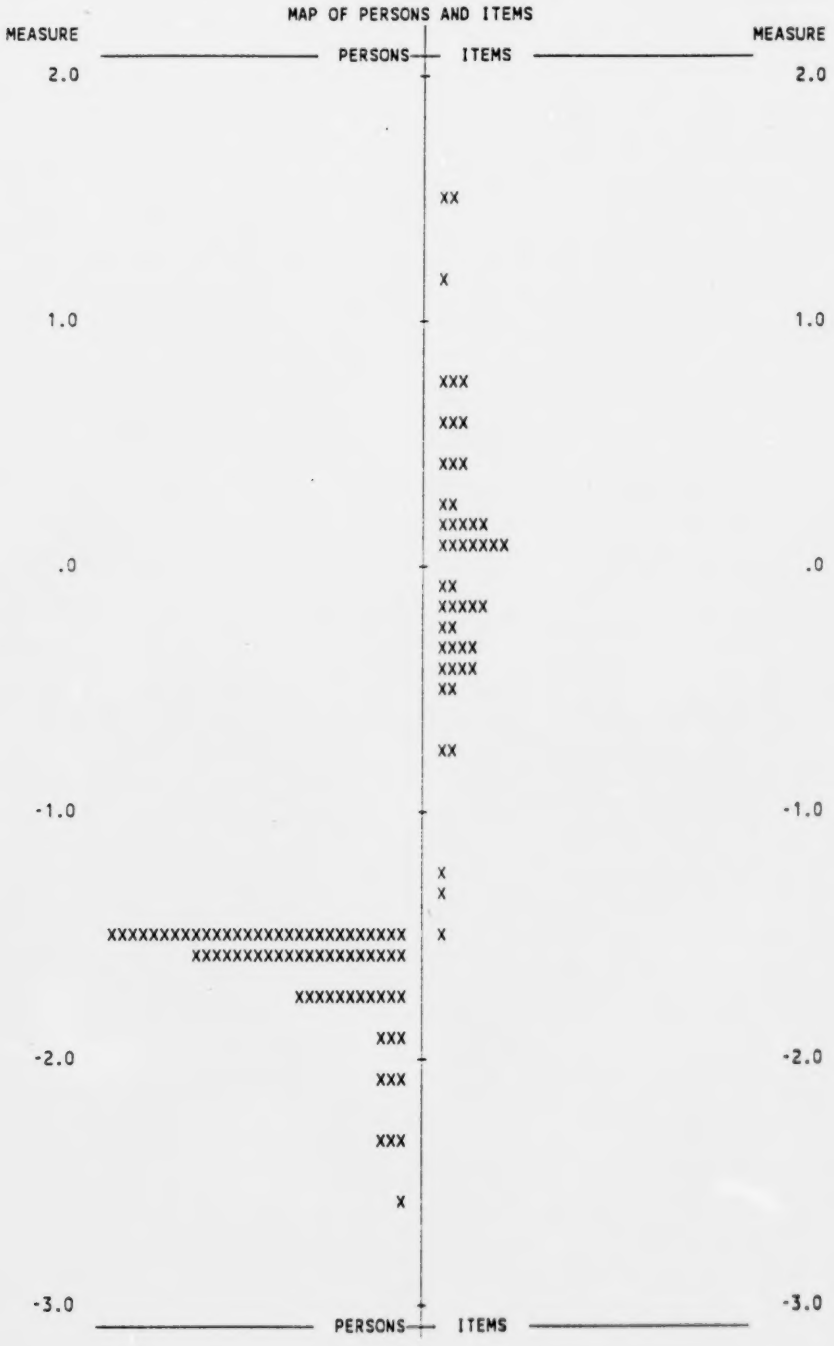
Wright, B.D. and Stone, M.H. (1979) *Best test design*. Chicago: MESA Press.

## APPENDIX

(Refer to Table 2, p 6. for a description of groups)

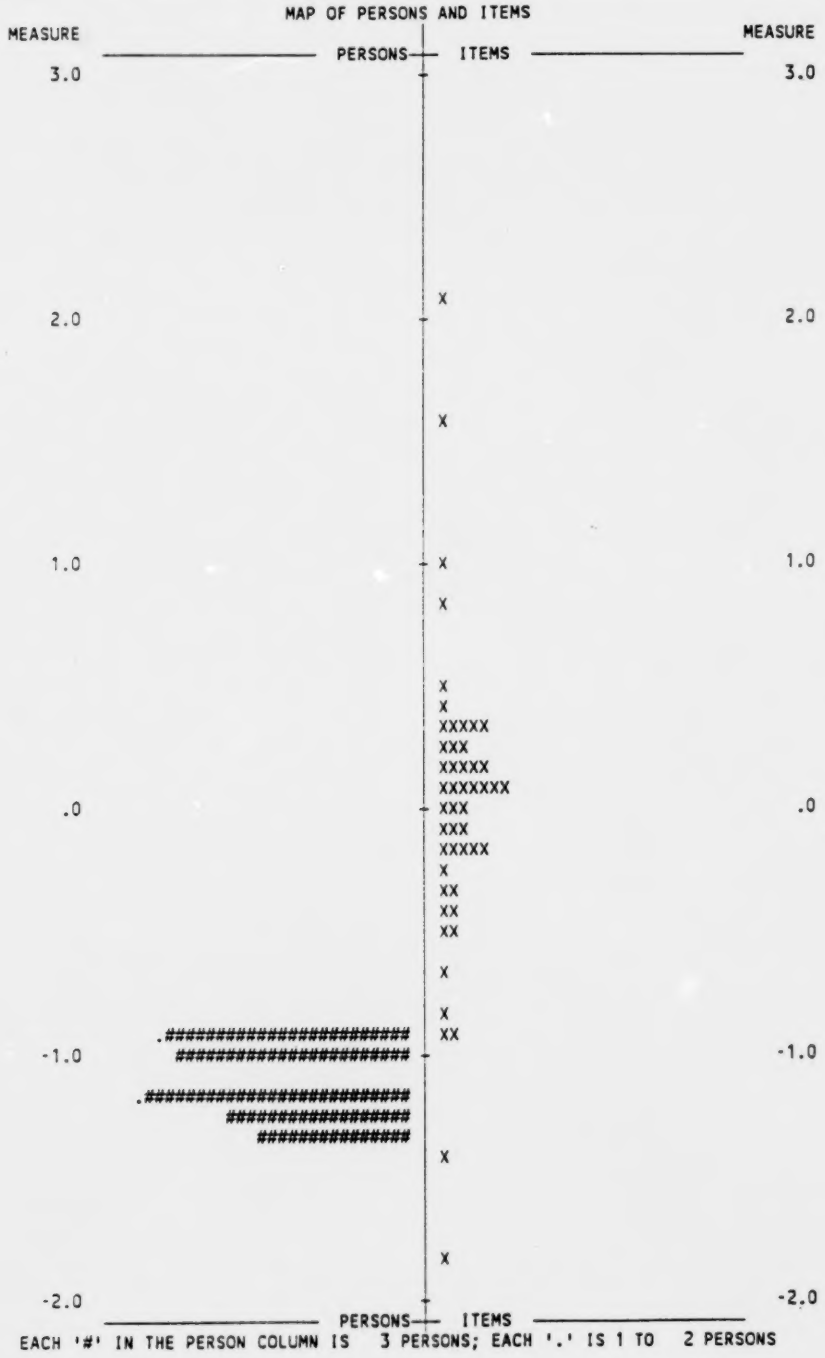
## GROUP 1

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 71 PERSONS 50 ITEMS



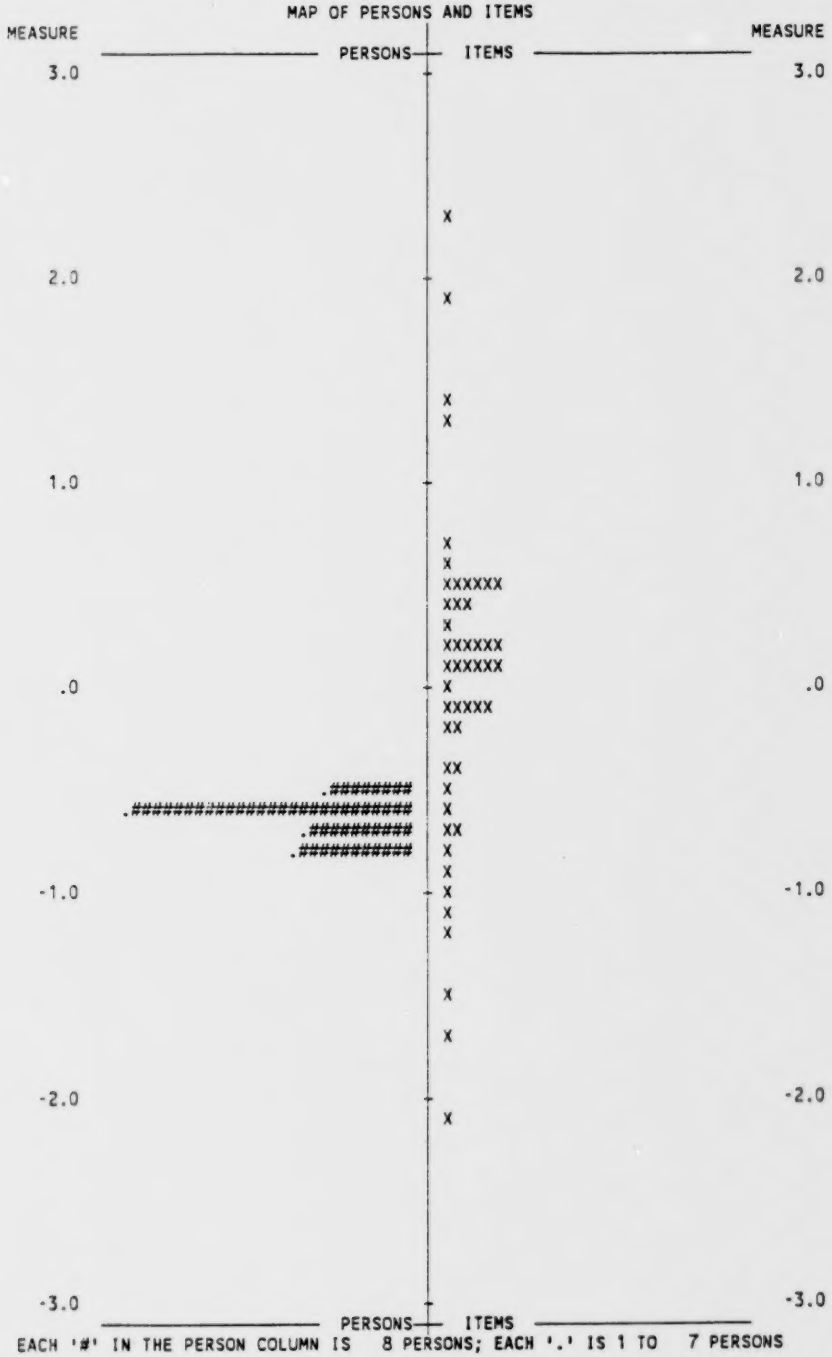
## GROUP 2

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 321 PERSONS 50 ITEMS



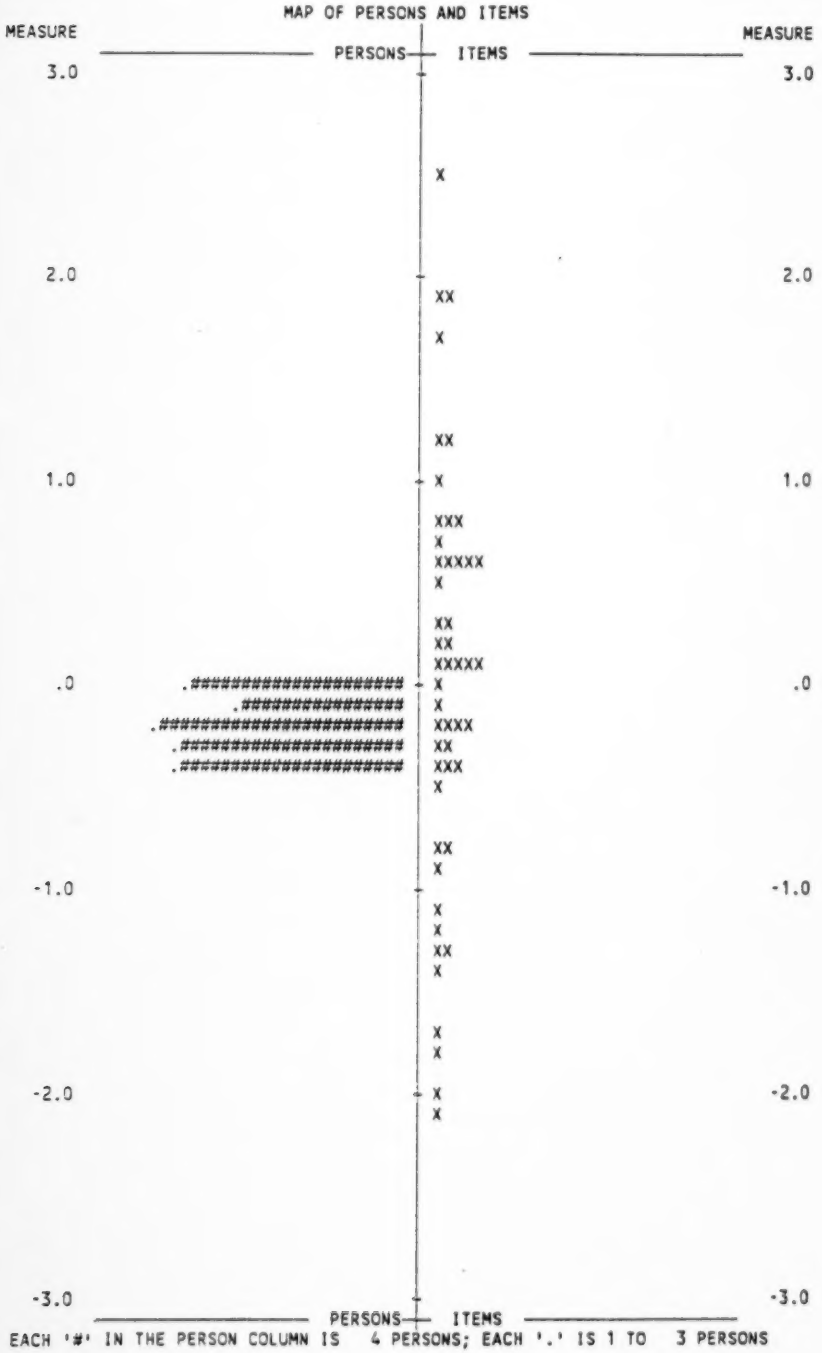
## GROUP 3

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 463 PERSONS 50 ITEMS



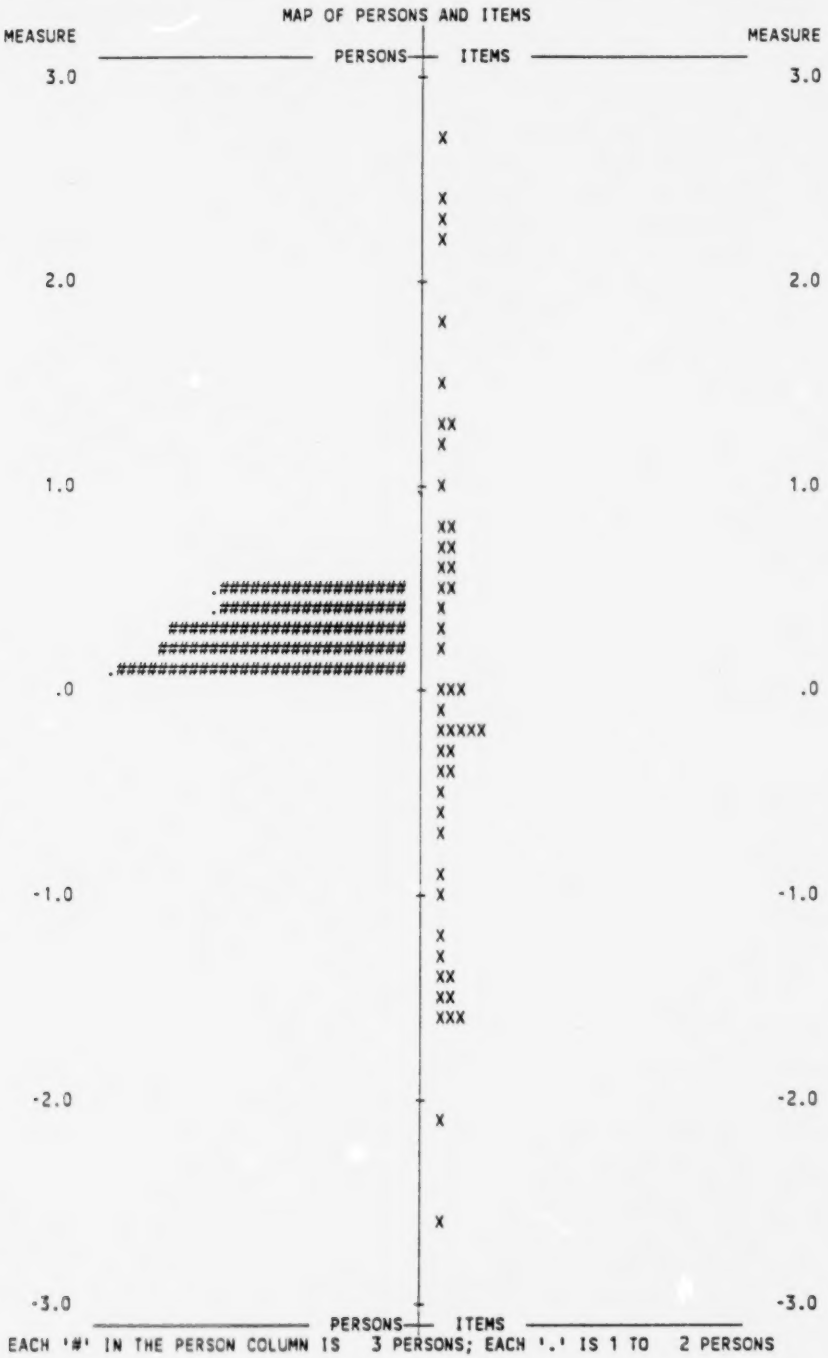
## GROUP 4

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 432 PERSONS 50 ITEMS



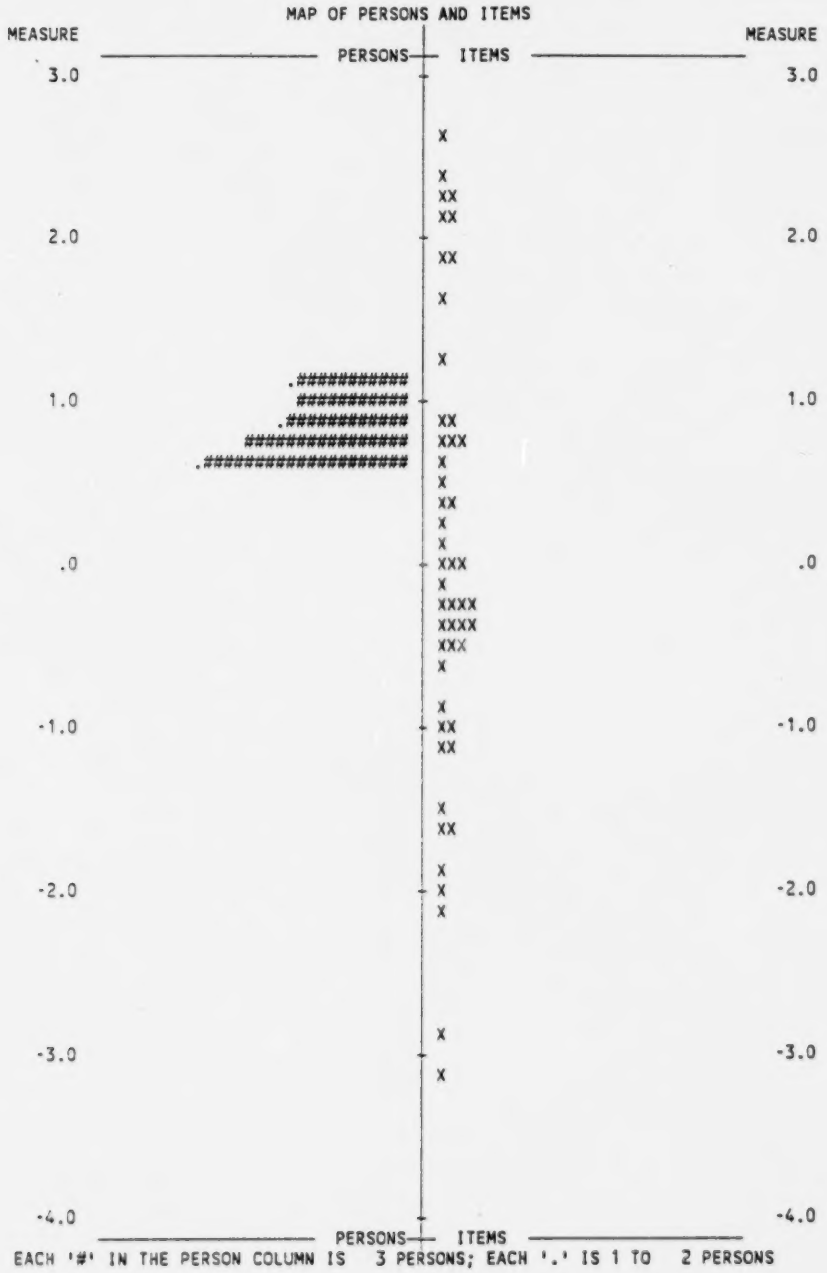
## GROUP 5

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 338 PERSONS 50 ITEMS



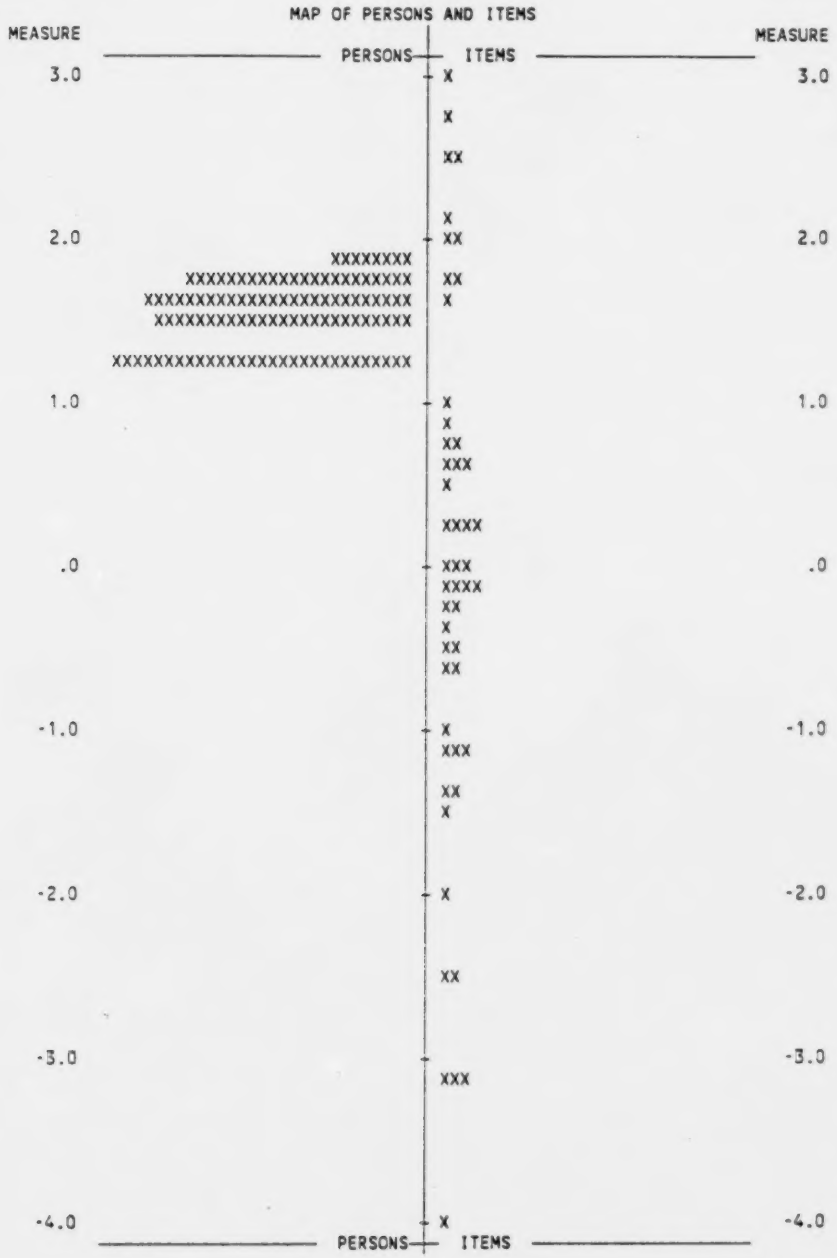
## GROUP 6

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 215 PERSONS 50 ITEMS



## GROUP 7

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 110 PERSONS 49 ITEMS



## GROUP 8

"BIGSTEPS" RASCH ANALYSIS VER. 2.25 ANALYZED: 67 PERSONS 42 ITEMS